TILBURG UNIVERSITY

Tilburg University

On the Harm that Pretesting Does

Danilov, D.L.; Magnus, J.R.

Publication date:
2001

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Danilov, D. L., & Magnus, J. R. (2001). *On the Harm that Pretesting Does*. (CentER Discussion Paper; Vol. 2001-37). Econometrics.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

No. 2001-37

ON THE HARM THAT PRETESTING DOES

By Dmitri L. Danilov and Jan R. Magnus

June 2001

ISSN 0924-7815

Discussion paper

On the harm that pretesting does *

Dmitri L. Danilov
and
Jan R. Magnus
CentER, Tilburg University

May 23, 2001

Affiliation: CentER, Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

phone: +31-13-466-3092
fax: +31-13-466-3066
email: magnus@kub.nl

*We are grateful to seminar participants at the University of Michigan, Michigan State University, the University of Wisconsin at Madison, and the Tinbergen Institute, University of Amsterdam for constructive and useful comments. Correspondence to: Jan R. Magnus, CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: magnus@kub.nl.

Title: On the harm that pretesting does

Proposed running head: On the harm that pretesting does

Mailing address:
Jan R. Magnus
CentER, Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

Keywords: Pretest estimator, Model selection, Mean squared error

JEL Codes: C13, C51

Abstract: Data in econometrics are, as a rule, non-experimental and hence we have to use the same data set to select the model and also to estimate the parameters in the selected model. In standard applied econometrics practice, however, one reports zero bias and some variance of the (pretest) estimators *conditional* on the selected model.

In this paper we find the *unconditional* moments of the pretest estimator, taking full account of the fact that model selection and estimation are an integrated procedure. We derive the bias, variance, and mean squared error of the pretest estimator, and show what the error is in not reporting the correct moments. This error can be very substantial. We also show that there can be large differences in underreporting between different model selection procedures. Finally, we ask how the underreporting error increases when the number of auxiliary regressors increases.

1 Introduction

In econometrics, due to the non-experimental nature of our discipline, the same data set is commonly used for model selection and for estimation. Standard statistical theory, as developed for the experimental sciences (biology, medicine, physics), is therefore not directly applicable, since the properties of most estimators in econometrics depend not only on the stochastic nature of the selected model, but also on the way this model was selected.

The simplest example of this situation is the standard linear model $y = X\beta + \gamma z + \varepsilon$, where we are uncertain whether to include z or not. The usual procedure is to compute the t-statistic on γ, and then, depending on whether $|t|$ is 'large' or 'small', decide to use the unrestricted or the restricted model. We then estimate β from the selected model. This estimator is a *pretest* estimator, but we commonly report its properties as if estimation had not been preceded by model selection. Thus we report no bias and an incorrect variance.

This is clearly wrong. Our view is *not* that we should avoid pretesting, even though it is well-known that pretest estimators have poor properties, inadmissibility being only one of them. This would be near-impossible in applied work.[1] Our view is simply that we should correctly report the bias and variance (or mean squared error) of the estimators, taking full account of the fact that model selection and estimation are an integrated procedure. This paper attempts to do this.

The literature on pretesting starts with Bancroft's (1944) famous article. Bancroft is mostly concerned with the bias introduced by pretests of homogeneity of variances and pretests of a regression coefficient. He considers the simplest case, in our notation $y = \beta x + \gamma z + \varepsilon$ (one β, one γ), where he wishes to estimate β while being uncertain about whether z should be in the regression or not. He then investigates the bias of the pretest estimator of β. Mosteller (1948) considers the special case $x' = (\imath', \imath')$, $z' = (0', \imath')$, where $\imath$ denotes the vector of ones. Thus, Mosteller considers pooling: if $\gamma = 0$ we pool, otherwise we don't pool. In this context, he calculates the mean squared error of the pretest estimator. Huntsberger (1955) extends

[1] There are, of course, Bayesian alternatives that avoid model selection. Judge and Bock (1978, 1983) provide a discussion of these. See also Zaman (1984).

Mosteller's paper by explicitly writing the pretest estimator as a (continuous) weighted average of the restricted ($\gamma = 0$) and unrestricted estimator, where the weights are functions of the relevant t-statistic. The fact that the pretest estimator has many undesirable properties is highlighted by Sclove, Morris and Radhakrishnan (1972). Feldstein (1973) is concerned with the problem of estimating β when x and z are highly correlated. He studies the pretest estimator and Huntsberger's weighted average estimator and obtains insights through a simulation experiment. The early literature is discussed in detail in Judge and Bock's (1978) important monograph.

Lovell (1983) asks what will be the true significance level of a t-test after pretesting, and recommends a simple rule-of-thumb. Roehrig (1984) establishes the relationship between the mean squared error of the pretest estimator and the mean squared error of the estimator of the nuisance parameters, a result later generalized by Magnus and Durbin (1999). Mittelhammer (1984) compares the risk functions of several estimators (including the pretest) under model misspecification, and concludes *inter alia* that all alternatives to OLS can be inferior to OLS in terms of prediction risk. The literature of this period is well summarized in Judge and Bock (1983) and in the special issue of the *Journal of Econometrics* (1984), edited by George Judge.

More recently, pretesting has attracted attention in finance, see for example Lo and MacKinlay (1990). Asymptotic aspects are considered in Sen (1979), Pötscher (1991), Zhang (1992), and Pötscher and Novak (1998). While most studies, including ours, are confined to the first two moments of the pretest statistics, Giles and Srivastava (1993) derive the distribution of the traditional pretest estimator. Summaries of the latest developments are given in Miller (1990), Giles and Giles (1993), Chatfield (1995), and Magnus (1999).

White (2000), building on work by Diebold and Mariano (1995) and West (1996), provides a method for testing the null hypothesis that the selected model has no predictive superiority over a benchmark model. Different model selection strategies (especially general-to-specific and specific-to-general) are discussed by Hoover and Perez (1999), who favor the general-to-specific procedure. Hendry (2001) advertises computer-automated general-to-specific procedures and claims that these procedures perform well in Monte Carlo experiments. We also find evidence that general-to-specific is preferable over

specific-to-general, and find the exact finite sample properties of the two procedures.

In spite of all this literature, we are still far removed from having a fully integrated procedure of model selection and parameter estimation. The current paper attempts to narrow this gap. Our main tool is a generalization of the 'Equivalence Theorem' of Magnus and Durbin (1999). We derive the bias, variance, and mean squared error of the pretest estimator, and show what the error is in not reporting the correct moments. This error can be very substantial. We also show that there can be large differences in underreporting between different model selection procedures. Finally, we ask how the underreporting error increases when the number of auxiliary regressors $z_1, \ldots, z_m$ increases.

The paper is organized as follows. We define the formal framework and the notation in Section 2. In Section 3 we prove two theorems, which form the basis of the subsequent analysis. Theorem 2 is a generalization of the 'Equivalence Theorem'. In Section 4 we discuss underreporting and its bounds. Section 5 discusses the simplest case, where there is only one auxiliary regressor z. There is only one possible pretest procedure here (using the t-statistic), and we find, among other things, that in the worst case we report only 13% of the actual pretest mean squared error. In Sections 6 and 7 we address the more difficult case where we have two auxiliary z regressors. Then, there is no unique selection procedure. We show, *inter alia*, that there can be large differences between general-to-specific and specific-to-general model selection. Section 8 briefly discusses various extensions and concludes the paper.

2 Set-up and notation

The set-up is the same as in Magnus and Durbin (1999) and is briefly summarized. We consider the standard linear regression model

$$y = X\beta + Z\gamma + \varepsilon \tag{1}$$

where $y\,(n \times 1)$ is the vector of observations, $X\,(n \times k)$ and $Z\,(n \times m)$ are matrices of nonrandom regressors, $\varepsilon\,(n \times 1)$ is a random vector of unobservable disturbances, and $\beta\,(k \times 1)$ and $\gamma\,(m \times 1)$ are unknown nonrandom parameter vectors. We assume that $k \geq 1$, $m \geq 1$, $n - k - m \geq 1$, that the design matrix

$(X : Z)$ has full column-rank $k + m$, and that the disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.

The reason for distinguishing between X and Z is that X contains explanatory variables that we want in the model on theoretical or other grounds (irrespective of the found t-values of the β-parameters), while Z contains additional explanatory variables of which we are less certain. Our focus is the estimation of β. Hence the only role for Z is to improve the estimation of β, while γ is a vector of nuisance parameters. The columns of X are called 'focus' regressors, and the columns of Z 'auxiliary' regressors.

We define the matrices

$$M = I_n - X(X'X)^{-1}X' \quad \text{and} \quad Q = (X'X)^{-1}X'Z(Z'MZ)^{-1/2},$$

and the scaled parameter vector $\eta = (Z'MZ)^{1/2}\gamma/\sigma$. The matrix Q can be interpreted as the (scaled) correlation between X and Z. Clearly, $Q = 0$ if and only if Z is orthogonal to X. The least-squares (LS) estimators of β and γ are $b_u = b_r - Q\hat{\theta}$ and $\hat{\gamma} = (Z'MZ)^{-1}Z'My$, where $b_r = (X'X)^{-1}X'y$ and $\hat{\theta} = (Z'MZ)^{1/2}\hat{\gamma}$. The subscripts '$u$' and '$r$' denote 'unrestricted' and 'restricted' (with $\gamma = 0$) respectively. Letting $\hat{\eta} = \hat{\theta}/\sigma$, we see that $\hat{\eta} \sim \mathrm{N}(\eta, I_m)$. Notice that $\hat{\eta}$ is only observable when σ is known, while $\hat{\theta}$ is observable whether σ is known or not.

3 The equivalence theorem generalized

Magnus and Durbin (1999) considered the estimation of β in model (1) and proposed a weighted-average least-squares (WALS) estimator of β of the form $b = \lambda b_u + (1 - \lambda)b_r$, where $\lambda = \lambda(\hat{\theta}, s_u^2)$ and s_u^2 denotes the estimator for σ^2 in the unrestricted model. This includes the usual pretest estimator as a special case, but only when one restricts the choice of model to the fully restricted and the fully unrestricted case. In this section we prove a generalization of the 'Equivalence Theorem' of Magnus and Durbin, which will allow us to consider not only the unrestricted estimator b_u and the restricted estimator b_r (where *all* γ's are set equal to zero), but also many or all intermediate estimators where *some* of the γ's are set equal to zero. We first state the following preliminary result.

Theorem 1: Let S_i be an $m \times r_i$ matrix of rank $r_i \geq 0$. The LS estimators of β and γ under the restriction $S_i'\gamma = 0$ are given by

$$b_{(i)} = b_r - QW_i\hat{\theta}, \qquad c_{(i)} = (Z'MZ)^{-1/2}W_i\hat{\theta},$$

where

$$W_i = I_m - P_i, \qquad P_i = (Z'MZ)^{-1/2}S_i\big(S_i'(Z'MZ)^{-1}S_i\big)^{-1}S_i'(Z'MZ)^{-1/2}$$

are symmetric idempotent $m \times m$ matrices of ranks $m - r_i$ and r_i respectively. (If $r_i = 0$ then $P_i = 0$.) The residual vector is

$$e_{(i)} = y - Xb_{(i)} - Zc_{(i)} = D_i y,$$

where

$$D_i = M - MZ(Z'MZ)^{-1/2}W_i(Z'MZ)^{-1/2}Z'M$$

is a symmetric idempotent matrix of rank $n - k - m + r_i$. The distribution of $b_{(i)}$ is given by

$$b_{(i)} \sim \mathrm{N}\left(\beta + \sigma QP_i\eta,\ \sigma^2\left((X'X)^{-1} + QW_iQ'\right)\right),$$

and the distribution of $s_{(i)}^2 = e_{(i)}'e_{(i)}/(n - k - m + r_i)$ by

$$\frac{(n - k - m + r_i)s_{(i)}^2}{\sigma^2} \sim \chi^2(n - k - m + r_i, \eta'P_i\eta).$$

Proof: Let $X_* = (X : Z)$, $\beta_*' = (\beta', \gamma')$, and $R = (0 : S_i')$. The estimator of β_* in the model $y = X_*\beta_* + u$ under the restriction $R\beta_* = 0$ is then given by

$$b_* = (X_*'X_*)^{-1}X_*'y - (X_*'X_*)^{-1}R'(R(X_*'X_*)^{-1}R')^{-1}R(X_*'X_*)^{-1}X_*'y.$$

Noting that

$$(X_*'X_*)^{-1} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} = \begin{pmatrix} (X'X)^{-1} + QQ' & -Q(Z'MZ)^{-1/2} \\ -(Z'MZ)^{-1/2}Q' & (Z'MZ)^{-1} \end{pmatrix},$$

and simplifying, the results follow. ‖

Several comments are in order. First, we should think of the matrix S_i as a selection matrix such as $S_i' = (0 : I_{r_i})$, although the theorem does not

depend on this. Secondly, if $Q = 0$ (that is, when Z is orthogonal to X) then $b_{(i)} = b_r$ whatever restriction is put on γ, but this is not so for $s^2_{(i)}$. In fact, $s^2_u \leq s^2_{(i)} \leq s^2_r$, where s^2_u and s^2_r denote the estimators for σ^2 in the unrestricted and restricted ($\gamma = 0$) models, respectively. Hence, if $Q = 0$, the pretest estimator is not affected by model selection, but its variance is (see also also footnote 3). Thirdly, the normality assumption plays a very minor role in Theorem 1. If we only assume that $\varepsilon \sim (0, \sigma^2 I_n)$, then the expressions for $b_{(i)}$ and $s^2_{(i)}$, the first two moments of $b_{(i)}$, and the first moment of $s^2_{(i)}$ remain the same. Finally, we notice that the partially restricted estimator $b_{(i)}$ is written as a linear function of two vectors b_r and $\hat{\theta}$, which are independent (since $X'y$ and $Z'My$ are independent).[2] Also, $c_{(i)}$ is a linear function of $\hat{\theta}$ only and hence independent of b_r.

If σ^2 is known, then any pretest procedure will use t- and F-statistics which depend on $\hat{\theta}$ only. If σ^2 is not known and estimated by s^2_u, then all t- and F-statistics will depend on $(\hat{\theta}, s^2_u)$. Now, it is a basic result in least-squares theory that s^2_u is independent of $(b_u, \hat{\gamma})$. It follows that b_r is independent of s^2_u. Hence, b_r will be independent of $(\hat{\theta}, s^2_u)$. Finally, if σ^2 is not known and estimated by $s^2_{(i)}$ corresponding to the selection matrix S_i, then it is no longer true that all t- and F-statistics depend only on $(\hat{\theta}, s^2_u)$. However, they still depend only on My, since both $c_{(i)}$ and $e_{(i)}$ are linear functions of My. Hence, the simple fact that b_r and $\hat{\theta}$ are independent implies that all t- and F-statistics used in a pretest procedure, and thus the choice of model, will be independent of b_r.

We are interested in WALS estimators of β, defined as

$$b = \sum_i \lambda_i b_{(i)}. \tag{2}$$

Motivated by the previous paragraph, we assume that the weights λ_i satisfy $\lambda_i = \lambda_i(My)$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Then,

$$b = b_r - QW\hat{\theta},$$

[2] In fact, even if the observations $y_1, \ldots, y_n$ are not normal and the data-generating process is unknown, b_r and $\hat{\theta}$ will still be uncorrelated, as long as the $\{y_i\}$ are uncorrelated with constant variance (Leeb and Pötscher (2000), Lemma A.1).

where

$$W = I_m - P, \qquad P = \sum_i \lambda_i P_i.$$

Notice that, while P_i and W_i are nonrandom matrices, P and W are random.

Theorem 2 (Equivalence theorem, generalized): Let $b = \sum_i \lambda_i b_{(i)}$, where $\lambda_i = \lambda_i(My)$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Then,

$$\mathrm{E}\, b = \beta - \sigma Q\, \mathrm{E}(W\hat{\eta} - \eta), \qquad \mathrm{var}(b) = \sigma^2 \left((X'X)^{-1} + Q\, \mathrm{var}(W\hat{\eta})Q'\right),$$

and hence

$$\mathrm{MSE}(b) = \sigma^2 \left((X'X)^{-1} + Q\, \mathrm{MSE}(W\hat{\eta})Q'\right).$$

Proof: Since b_r and My are independent, we have

$$\mathrm{E}(b_r \mid My) = \mathrm{E}(b_r), \qquad \mathrm{var}(b_r \mid My) = \mathrm{var}(b_r).$$

Hence,

$$\begin{aligned} \mathrm{E}(b \mid My) &= \mathrm{E}(b_r \mid My) - Q\, \mathrm{E}(W\hat{\theta} \mid My) \\ &= \mathrm{E}(b_r) - \sigma Q W\hat{\eta} = \beta - \sigma Q(W\hat{\eta} - \eta) \end{aligned}$$

and

$$\mathrm{var}(b \mid My) = \mathrm{var}(b_r \mid My) \ = \mathrm{var}(b_r) = \sigma^2 (X'X)^{-1}.$$

The unconditional mean and variance of b and hence its mean squared error follow. ‖

This provides a nontrivial generalization, using a simpler proof, of Theorem 2 in Magnus and Durbin (1999). Apparently, the properties of the complicated pretest estimator b of β depend critically on the properties of the less complicated estimator $W\hat{\eta}$ of η.

The restriction that λ_i must depend only on My is a very light one. This allows not only all standard pretest procedures, but also inequality-constrained least squares. Thus, Theorem 2 explains the 'surprising symmetry' found by Thomson and Schmidt (1982, p. 176). The normality assumption plays a stronger role in Theorem 2 than in Theorem 1. Still, if we only assume that $\varepsilon \sim (0, \sigma^2 I_n)$, then Theorem 2 will still hold if the mean and variance of b_r conditional on My are equal to the unconditional mean and variance of b_r.

4 Pretesting and underreporting

Theorem 2 shows that if we can find λ_i's such that $W\hat{\eta}$ is an optimal estimator of η, then the same λ_i's will provide an optimal WALS estimator of β. In this paper, however, we are not interested in finding λ_i's such that $W\hat{\eta}$ is an optimal estimator of η. Instead we are interested in the commonly used pretest estimator.

In the idealized context of the linear model $y = X\beta + Z\gamma + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, we define a *pretest procedure* as a two-step procedure. In step 1 we select the model. In the case $m = 1$ there are two models to choose from: the unrestricted and the restricted (where $\gamma = 0$). In the case $m = 2$ there are four possible models: the unrestricted model, two partially restricted models (one of the two γ's is zero), and the restricted model (both γ's are zero). In general, there are 2^m models to consider in a pretest procedure. We require that the model selection criterion depends on y only through My. In step 2 we estimate the unknown parameters β (and σ^2) from the selected model. This yields the pretest estimators b (and s^2). In a pretest procedure thus defined, the selection matrices take the form $S_i' = (I_{r_i} : 0)$ or a column-permutation thereof and the λ_i's are all zero except one which is one.

The mean squared error of the pretest estimator b is, according to Theorem 2,

$$\mathrm{MSE}(b) = \sigma^2 \left((X'X)^{-1} + Q \,\mathrm{MSE}(W\hat{\eta}) Q' \right).$$

In applied econometrics practice the same estimator b is selected, but the the effects of pretesting are ignored, the reported bias is zero, and hence the reported MSE equals the reported variance. If we assume that σ^2 is known, then the reported MSE equals

$$\widetilde{\mathrm{MSE}}(b) = \sigma^2 \left((X'X)^{-1} + QWQ' \right),$$

according to Theorem 1, since $W = W_i$ if the i-th model is selected. Notice that $\widetilde{\mathrm{MSE}}(b)$ is random since W is random. Let $\omega'\beta$ be our focus parameter, where ω is an arbitrary nonzero $k \times 1$ vector. In order to compare

$$\mathrm{MSE}(\omega' b) = \sigma^2 \left(\omega'(X'X)^{-1}\omega + \omega' Q \,\mathrm{MSE}(W\hat{\eta}) Q'\omega \right) \tag{3}$$

with

$$\widetilde{\mathrm{MSE}}(\omega' b) = \sigma^2 \left(\omega'(X'X)^{-1}\omega + \omega' QWQ'\omega\right), \tag{4}$$

we define the *underreporting ratio* UR as one minus the ratio of (4) and (3). Thus,

$$\mathrm{UR} = 1 - \frac{\widetilde{\mathrm{MSE}}(\omega' b)}{\mathrm{MSE}(\omega' b)} = \frac{q'(R(\eta) - W)q}{q'R(\eta)q + (1/q_0^2)}, \tag{5}$$

where

$$R(\eta) = \mathrm{MSE}(W\hat{\eta}), \qquad q = \frac{Q'\omega}{\sqrt{\omega' QQ'\omega}}, \qquad q_0^2 = \frac{\omega' QQ'\omega}{\omega'(X'X)^{-1}\omega}.$$

Notice that $q'q = 1$. The UR is a random variable, since it depends on W, which depends on $\hat{\eta}$. Both the UR and its expectation are unobservable, since they depend on η via $R(\eta)$.

One would expect that the matrix $\mathrm{MSE}(b)$ is at least as large as the matrix $\mathrm{E}(\widetilde{\mathrm{MSE}}(b))$ (in the sense that their difference is positive semidefinite), because pretesting introduces additional noise which is ignored in the reported MSE. Since

$$\mathrm{MSE}(W\hat{\eta}) = \sum_{i=1}^{2^m} \mathrm{E}\,\lambda_i (W_i\hat{\eta} - \eta)(W_i\hat{\eta} - \eta)'$$

and

$$\mathrm{E}(W) = \sum_{i=1}^{2^m} (\mathrm{E}\,\lambda_i) W_i,$$

this is guaranteed if the matrix

$$\sum_{i=1}^{2^m} \mathrm{E}\,\lambda_i \left((W_i\hat{\eta} - \eta)(W_i\hat{\eta} - \eta)' - W_i\right) \tag{6}$$

is positive semidefinite. We shall see in the next section that it is possible to devise pretest procedures which do not satisfy this requirement. Such procedures, however, tend to be rather silly. We shall say that a pretest procedure is *viable* if the matrix in (6) is positive semidefinite over the whole parameter space. For any viable pretest procedure, E(UR) is a number between zero

and one. When q_0^2 (known to the investigator) tends to zero, then there is no underreporting: $\mathrm{E(UR)} = 0$.[3] But when q_0^2 is large, E(UR) can be close to one.

The $m \times m$ matrix $\mathrm{E}(W)$ is a weighted average of idempotent matrices, and hence is bounded: all its elements are ≤ 1 in absolute value, and all its diagonal elements (and all its eigenvalues) lie in the interval $[0, 1]$. In fact,

$$0 \leq \pi_u \leq \xi_j(\mathrm{E}\, W) \leq 1 - \pi_r \leq 1 \qquad (j = 1, \ldots, m),$$

where $\xi_j(A)$ denotes the j-th eigenvalue of A, π_u is the probability of choosing the unrestricted model ($P_i = 0$), and π_r the probability of choosing the restricted model ($P_i = I_m$).

The E(UR) is a function of q (normalized by $q'q = 1$), q_0^2, η, and $Z'MZ$ (and m). Maximizing over q gives the inequality

$$\mathrm{E(UR)} \leq q_0^2 \max_{1 \leq j \leq m} \xi_j \left((I_m + q_0^2 R)^{-1/2} (R - \mathrm{E}\, W)(I_m + q_0^2 R)^{-1/2} \right), \qquad (7)$$

where $R = \mathrm{MSE}(W\hat{\eta})$. Then, letting

$$\mathrm{E}^*(\mathrm{UR}) = \max_{q, q_0^2} \mathrm{E(UR)},$$

we find, as $q_0^2 \to \infty$,

$$\mathrm{E}^*(\mathrm{UR}) = 1 - \min_{1 \leq j \leq m} \xi_j(R^{-1/2}(\mathrm{E}\, W)R^{-1/2}) \leq 1 - \frac{\pi_u}{\max_j \xi_j(R)}, \qquad (8)$$

which depends on η and $Z'MZ$ (and m). We see from (8) that the expected UR can be arbitrarily close to 1 if the mean squared error R fails to be bounded in η. This can not happen when $m = 1$ (unless we *always* choose the restricted model, whatever the value of the observed t-statistic), but it can happen when $m \geq 2$, as we shall see in Section 7.

Finally, since E(UR) depends on $Z'MZ$, we briefly consider the role of this matrix. Without loss of generality, we may scale all z variables so that $z_j' M z_j = 1$ for all $j = 1, \ldots, m$. In the special case where we can choose the z variables to be 'orthogonal' (in the sense that Mz_i and Mz_j are orthogonal for every $i \neq j$), we have $Z'MZ = I_m$, and major simplifications occur.

[3] This happens when $X'Z \to 0$, but also (more generally and less trivially) when $Q'\omega = 0$. In either case $b = b_r$ whatever pretesting we do.

Theorem 3: Assume that the weights $\lambda_i = \lambda_i(My)$, $i = 1, \ldots, 2^m$, are defined by some pretest procedure, so that they are all zero except one which is one. Let $\lambda(x) = 1$ if $|x| > c$ for some $c > 0$, and 0 otherwise. In the special case $Z'MZ = I_m$:

a. W is a diagonal matrix with typical element $w_{jj} = \lambda(\hat{\eta}_j)$;

b. $\mathrm{MSE}(W\hat{\eta}) = V + dd'$, where V is a diagonal $m \times m$ matrix and d an $m \times 1$ vector with typical elements

$$v_{jj} = \mathrm{var}(\lambda(\hat{\eta}_j)\hat{\eta}_j), \qquad d_j = \mathrm{E}(\lambda(\hat{\eta}_j)\hat{\eta}_j - \eta_j);$$

c. The decision whether or not to include z_j in the regression is based exclusively on the t-statistic $\hat{\eta}_j$, and is independent of the selection procedure.

Proof: Using Theorem 1, we have $P_i = S_i(S_i'S_i)^{-1}S_i'$, and, since S_i' is a selection matrix of the form $(I_{r_i} : 0)$ or a column-permutation thereof, it follows that $S_i'S_i = I_{r_i}$ and hence that P_i is a diagonal matrix with r_i ones and $m - r_i$ zeros on the diagonal, and that W_i is a diagonal matrix with $m - r_i$ ones and r_i zeros on the diagonal. Now, also by Theorem 1, $c_{(i)} = W_i\hat{\theta}$ is the estimator of γ under the restriction $S_i'\gamma = 0$. Hence, the estimator of γ_j under this restriction is the j-th component of $c_{(i)}$, which is either 0 (if z_j is excluded from the model) or $\hat{\theta}_j$ (if z_j is included). Thus all models which include z_j as a regressor will have the *same* estimator of γ_j, irrespective which other γ's are estimated. This implies c. Clearly, W is diagonal. The j-th diagonal element w_{jj} is either 0 (if z_j is excluded from the model) or 1 (if z_j is included), that is, $w_{jj} = \lambda(\hat{\eta}_j)$. This implies a. It also implies that the components of $W\hat{\eta}$ are independent of each other, and hence b. follows. ||

Since we shall see that the choice of model selection procedure may matter a lot for the properties of the estimated focus parameters, it is advisable — if at all possible — to choose the auxiliary regressors such that $Z'MZ = I_m$. This will not only make the the pretest estimator independent of the chosen model selection procedure, but it also allows us to obtain explicit analytical expressions for the moments of the estimator, and it guarantees bounded risk

for any value of m. (In the general non-orthogonal case, risk is bounded for $m = 1$, but not necessarily for $m \geq 2$, see Section 7.)

5 Underreporting with one nuisance parameter

In the case of one nuisance parameter, the model becomes $y = X\beta + \gamma z + \varepsilon$, where the nuisance parameter γ is a scalar. We have only two models to compare: the unrestricted ($W_1 = 1$, $b_{(1)} = b_u$, $\lambda_1 = \lambda$) and the restricted ($W_2 = 0$, $b_{(2)} = b_r$, $\lambda_2 = 1 - \lambda$). As a result we find

$$b = \lambda b_u + (1 - \lambda) b_r, \qquad W = \lambda,$$

and

$$\mathrm{MSE}(W\hat{\eta}) = \mathrm{MSE}(\lambda\hat{\eta}) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \qquad \mathrm{E}\, W = \mathrm{E}\, \lambda.$$

The underreporting ratio is thus

$$\mathrm{UR}(\hat{\eta}, \eta) = \frac{R(\eta) - \lambda(\hat{\eta})}{R(\eta) + (1/q_0^2)},$$

where $\lambda(\hat{\eta}) = 1$ if $|\hat{\eta}| > c$ for some $c > 0$, and 0 otherwise, and

$$R(\eta) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \qquad q_0^2 = \frac{(z'X(X'X)^{-1}\omega)^2}{(z'Mz)(\omega'(X'X)^{-1}\omega)}.$$

Assuming again that σ^2 is known and that c is given (say, $c = 1.96$), the λ-function depends only on $\hat{\eta}$, R depends only on η, and hence the UR depends on q_0^2 and $\hat{\eta}$ (both known to the investigator), and η (unknown).

It is easy to see that the larger is $R(\eta)$, the larger is UR. The random variable $\lambda\hat{\eta}$, considered as an estimator of η, thus plays a crucial role in determining the amount of underreporting. We consider its squared bias, variance and MSE in Figure 1.

FIGURE 1

The bias of $\lambda\hat{\eta}$ is negative for $\eta > 0$ and reaches its minimum -0.66 at $\eta = 1.46$. The variance reaches its minimum 0.28 at $\eta = 0$ and its maximum 2.23 at $\eta = 2.34$. The MSE $R(\eta)$ is shaped similarly to the variance. It

reaches its minimum at $\eta = 0$ and its maximum 2.46 at $\eta = 2.16$. The variance of $\lambda\hat{\eta}$ is large relative to its bias, suggesting that variance-reduction is more important than bias-reduction.

We also graph the expectation of the reported MSE of $\lambda\hat{\eta}$, that is $\mathrm{E}(\lambda)$, as a function of η for $c = 1.96$, and the MSE of the unrestricted estimator of η, that is $\mathrm{MSE}(\hat{\eta})$ (the dashed line, constant at 1). Since λ only takes the values 0 and 1, $\mathrm{E}(\lambda)$ denotes the probability of choosing the unrestricted model ($\lambda = 1$). But λ also denotes the reported variance (MSE). We see that $\mathrm{E}(\lambda) \equiv \Pr(|\hat{\eta}| > c)$ increases monotonically between 0.05 at $\eta = 0$ and 1 at $\eta = \infty$. Since $\mathrm{MSE}(\lambda\hat{\eta}) \geq \mathrm{E}(\lambda)$, the pretest procedure is viable.[4]

Since λ can only take the values 0 and 1, we can graph the UR for these two values, together with the expected UR and the expectation of λ. This is done in Figure 2 for the case $q_0^2 = 1$.

FIGURE 2

Figure 2 contains four graphs: the UR at $\lambda = 1$ and at $\lambda = 0$, the expected UR, and $\mathrm{E}(\lambda)$. The graph labeled $\mathrm{UR}(\lambda = 0)$ gives the underreporting ratio when the restricted model is chosen. This function reaches its minimum 0.22 at $\eta = 0$, its maximum 0.71 at $\eta = 2.16$, and approaches $q_0^2/(1 + q_0^2) = 0.5$ as $\eta \to \infty$. Hence, for large values of η, only one half of the actual MSE will be reported when the restricted model is chosen.

Similarly, the graph $\mathrm{UR}(\lambda = 1)$ gives the underreporting ratio when the unrestricted model is chosen. It reaches its minimum -0.56 at $\eta = 0$, its maximum 0.42 at $\eta = 2.16$, and approaches 0 as $\eta \to \infty$. Thus, when η is large and we (correctly) choose the unrestricted model, the UR is zero (no underreporting), but when η is small and we (correctly) choose the restricted model, the UR is still 0.22.

Note that both $\mathrm{UR}(\lambda = 1)$ and $\mathrm{UR}(\lambda = 0)$ reach their maximum at $\eta = 2.16$, where also $\mathrm{MSE}(\lambda\hat{\eta})$ reaches its maximum. Moreover, the value 2.16 does not depend on q_0^2 (although it does depend on c). Note also that $\mathrm{UR}(\lambda = 1)$ is always smaller than $\mathrm{UR}(\lambda = 0)$, and hence that underreporting is higher if the restricted model is chosen.

[4]However, not all λ-functions lead to a viable procedure. For example, the — admittedly silly — procedure defined by $\lambda = 1$ if $|\hat{\eta}| \leq c$ and 0 otherwise is not viable, since $\mathrm{MSE}(\lambda\hat{\eta}) < \mathrm{E}(\lambda)$ at $\eta = 0$ for any $c > 0$.

When $\lambda = 0$ (and when consequently the restricted model is chosen), the UR always lies between 0 and 1. But when $\lambda = 1$ (unrestricted model), the UR can become negative. This occurs when $|\hat{\eta}|$ is large (> 1.96) but $|\eta|$ is small (< 0.84). In that case the reported MSE is larger than the pretest MSE. The probability that this happens (given by $\mathrm{E}(\lambda)$) is, however, small.

The underreporting ratio $\mathrm{UR}(\lambda = 1)$ does not take account of the probability that the event $\{\lambda = 1\}$ occurs. Neither does $\mathrm{UR}(\lambda = 0)$ take account of the probability that the event $\{\lambda = 0\}$ occurs. In contrast, the expected UR takes account of both probabilities, since it is a weighted average of $\mathrm{UR}(\lambda = 1)$ and $\mathrm{UR}(\lambda = 0)$ with weights $\mathrm{E}(\lambda)$ and $1 - \mathrm{E}(\lambda)$, respectively. We see that E(UR) is 0.18 at $\eta = 0$, reaches a maximum 0.57 at $\eta = 1.73$, and approaches the curve of $\mathrm{UR}(\lambda = 1)$ as η increases. The E(UR) varies substantially with η (from 0 to 0.57), indicating that on average the pretest MSE can be 2.3 times the reported MSE ($1/(1 - 0.57) = 2.3$). In contrast to the UR at $\lambda = 0$ or 1, the maximum of E(UR) *does* depend on q_0. This dependence is analyzed in Figure 3.

FIGURE 3

In Figure 3 we graph E(UR) for five different values of q_0^2: 0, 0.1, 1, 10, and ∞. At $q_0^2 = 0$ there is no underreporting and $\mathrm{E(UR)} = 0$. At $q_0^2 = \infty$, E(UR) is large; the maximum occurs at $\eta = 0.82$ where $\mathrm{E(UR)} = 0.87$. This means that the reported variance should be multiplied by about 7.5 in order to obtain the true MSE of the pretest estimator.

Finally, since both UR and E(UR) depend on η, we also consider the behavior of the underreporting ratio at $\eta = 1$. This is an interesting value, because it is the value of η where the investigator is indifferent between the restricted and the unrestricted model; see Magnus and Durbin (1999, Theorem 1).

FIGURE 4

Figure 4 shows that the UR at $\eta = 1$ is an increasing function of q_0^2, with $\mathrm{UR} = 0$ at $q_0^2 = 0$. When $q_0^2 \to \infty$, UR approaches 1 when $\lambda = 0$ and 0.20 when $\lambda = 1$, since $R(1) = 1.26$. The expectation of UR approaches 0.86, since $\mathrm{E}(\lambda) = 0.17$ at $\eta = 1$.

We conclude that the effect of not reporting the true bias and variance of the pretest estimator can lead to serious misrepresentation of the results, *even*

in the case $m = 1$. The larger is q_0^2 (known to the investigator), the larger will be the expected UR. For given q_0^2 we can draw the expected UR as a function of η, as in Figure 3, and calculate the maximum E(UR). Alternatively, we can calculate E(UR) at the point $\eta = \hat{\eta}$ and use this as an estimate of the seriousness of underreporting. The E(UR) can be as large as 0.87 (at $q_0^2 = \infty$ and $\eta = 0.82$). This means that in the worst case the expectation of the reported variance of the pretest estimator is only 13% of its actual mean squared error.

6 Model selection: general-to-specific and specific-to-general

When $m = 1$ pretesting is simple: look at the t-statistic for γ in the unrestricted model. If $|t| > c$, choose the unrestricted model (leading to b_u); otherwise choose the restricted model (leading to b_r). When $m > 1$ there are many ways to pretest. We consider the case $m = 2$ under the following conditions: model selection is based on t-statistics only, in the selected model all t-statistics are 'significant', and σ^2 is known.

Without loss of generality we normalize z_1 and z_2, the regressors associated with the nuisance parameters γ_1 and γ_2, by setting $z_i' M z_i = 1$ for $i = 1, 2$. Then,

$$Z'MZ = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where $|r| < 1$, and

$$(Z'MZ)^{-1/2} = \frac{1}{\sqrt{1-r^2}} \begin{pmatrix} \alpha & -\rho \\ -\rho & \alpha \end{pmatrix},$$

with

$$\alpha = \frac{\sqrt{1+r} + \sqrt{1-r}}{2}, \qquad \rho = \frac{\sqrt{1+r} - \sqrt{1-r}}{2}.$$

There are four t-statistics to consider: two in the unrestricted model (denoted t_1 and t_2), one in the model where $\gamma_2 = 0$ (denoted $t_{(1)}$), and one in the model where $\gamma_1 = 0$ (denoted $t_{(2)}$). Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the components of $\hat{\eta}$. Then, each of the four t-statistics is a linear function of $\hat{\eta}_1$ and $\hat{\eta}_2$ in accordance

with Theorem 1:

$$t_1 = \alpha\hat{\eta}_1 - \rho\hat{\eta}_2, \qquad t_2 = -\rho\hat{\eta}_1 + \alpha\hat{\eta}_2,$$

and

$$t_{(1)} = \alpha\hat{\eta}_1 + \rho\hat{\eta}_2, \qquad t_{(2)} = \rho\hat{\eta}_1 + \alpha\hat{\eta}_2.$$

Of course, since $\alpha^2 + \rho^2 = 1$, all four t-statistics are normally distributed with unit variance and, under the appropriate null hypothesis, mean zero. Also, $t_{(1)}$ is independent of t_2 and $t_{(2)}$ is independent of t_1, for the same reason that b_r and $\hat{\eta}$ are independent. Further,

$$\text{corr}(t_1, t_{(1)}) = \text{corr}(t_2, t_{(2)}) = \sqrt{1 - r^2} > 0,$$

and

$$\text{corr}(t_1, t_2) = -r, \qquad \text{corr}(t_{(1)}, t_{(2)}) = r.$$

Finally,

$$|t_1| > |t_2| \iff |t_{(1)}| > |t_{(2)}| \iff |\hat{\eta}_1| > |\hat{\eta}_2|.$$

A t-statistic is 'significant' if its absolute value exceeds some a priori chosen positive constant c, such as 1.96.

We shall investigate two pretest procedures that are in common use: 'general-to-specific' and 'specific-to-general'. Let $\mathcal{M}_0$ denote the restricted model, $\mathcal{M}_1$ the model with only z_1 ($\gamma_2 = 0$), $\mathcal{M}_2$ the model with only z_2 ($\gamma_1 = 0$), and $\mathcal{M}_{12}$ the unrestricted model. Then we define the general-to-specific (or 'backward' or 'top-down') procedure as follows:

a. Estimate the unrestricted model $\mathcal{M}_{12}$. This yields t-statistics t_1 and t_2;

b. Choose $\mathcal{M}_{12}$ if both t_1 and t_2 are significant;

c. Otherwise,

   (i) if $|t_1| > |t_2|$ estimate $\mathcal{M}_1$, yielding $t_{(1)}$. If $t_{(1)}$ is significant choose $\mathcal{M}_1$, otherwise choose $\mathcal{M}_0$;

   (ii) if $|t_1| \leq |t_2|$ estimate $\mathcal{M}_2$, yielding $t_{(2)}$. If $t_{(2)}$ is significant choose $\mathcal{M}_2$, otherwise choose $\mathcal{M}_0$.

Similarly, we define the specific-to-general (or 'forward' or 'bottom-up') procedure as follows:

a. Estimate both partially restricted models $\mathcal{M}_1$ and $\mathcal{M}_2$. This yields t-statistics $t_{(1)}$ and $t_{(2)}$;

b. Choose $\mathcal{M}_0$ if neither $t_{(1)}$ nor $t_{(2)}$ is significant;

c. Otherwise, estimate the unrestricted model yielding t_1 and t_2, and choose $\mathcal{M}_{12}$ if t_1 and t_2 are both significant;

d. In all other cases choose $\mathcal{M}_1$ (if $|t_{(1)}| > |t_{(2)}|$) or $\mathcal{M}_2$ (if $|t_{(1)}| \leq |t_{(2)}|$).

For $r = 0.8$, we graph the relevant regions in $(\hat{\eta}_1, \hat{\eta}_2)$-plane for both procedures in Figures 5 and 6.

FIGURES 5 AND 6

Since the two cases ($|t_{(1)}| \leq c < |t_1|$, $|t_2| \leq c < |t_{(2)}|$) and ($|t_{(2)}| \leq c < |t_2|$, $|t_1| \leq c < |t_{(1)}|$) can not occur, we see that both procedures are identical, except for the case where t_1 and t_2 are both significant, while $t_{(1)}$ and $t_{(2)}$ are both not significant. In that case, the general-to-specific procedure chooses the unrestricted model and the specific-to-general procedure chooses the restricted model. In the special case $r = 0$, we find $t_1 = t_{(1)} = \hat{\eta}_1$ and $t_2 = t_{(2)} = \hat{\eta}_2$, and all pretest procedures coincide. When $|r| \to 1$, the difference between the two procedures is at its largest. In spite of the seemingly small difference between the two pretest procedures, the effect of pretesting on underreporting will be surprisingly different for the two procedures.

7 Underreporting with two nuisance parameters

In the case $m = 1$ the expected underreporting ratio E(UR) depends (for fixed c) on two parameters: q_0^2 (known to the investigator) and η (unknown). In the case $m = 2$, E(UR) depends, after normalization, on five parameters: q_0^2, q_1 and r (known), and η_1 and η_2 (unknown). In addition, E(UR) depends on the procedure.

We have four models to compare: the unrestricted $\mathcal{M}_{12}$, the partially restricted $\mathcal{M}_1$ ($\gamma_2 = 0$) and $\mathcal{M}_2$ ($\gamma_1 = 0$), and the restricted $\mathcal{M}_0$ ($\gamma_1 = \gamma_2 = 0$). This implies selection matrices $S_0 = I_2$, $S_1 = (0,1)'$, and $S_2 = (1,0)'$ (The matrix S_{12} has no columns), and hence $W_0 = 0$, $W_{12} = I_2$,

$$W_1 = \frac{1}{2}\begin{pmatrix} 1+\sqrt{1-r^2} & r \\ r & 1-\sqrt{1-r^2} \end{pmatrix},$$

and

$$W_2 = \frac{1}{2}\begin{pmatrix} 1-\sqrt{1-r^2} & r \\ r & 1+\sqrt{1-r^2} \end{pmatrix}.$$

Since $W = \lambda_0 W_0 + \lambda_1 W_1 + \lambda_2 W_2 + \lambda_{12} W_{12}$, we thus find

$$W = \frac{1}{2}\begin{pmatrix} \mathrm{tr}(W) + \sqrt{1-r^2}(\lambda_1 - \lambda_2) & r(\lambda_1+\lambda_2) \\ r(\lambda_1+\lambda_2) & \mathrm{tr}(W) - \sqrt{1-r^2}(\lambda_1 - \lambda_2) \end{pmatrix},$$

where $\mathrm{tr}(W) = \lambda_1 + \lambda_2 + 2\lambda_{12}$. As before, let $\lambda(x) = 1$ if $|x| > c$ and 0 otherwise. Then,

$$\begin{aligned} \lambda_0 &= (1-\lambda(t_{(1)}))(1-\lambda(t_{(2)})) - \delta B_1, & \lambda_1 &= \lambda(t_{(1)})(1-\lambda(t_2)) - (1-\mu)B_2, \\ \lambda_2 &= \lambda(t_{(2)})(1-\lambda(t_1)) - \mu B_2, & \lambda_{12} &= \lambda(t_1)\lambda(t_2) - (1-\delta)B_1, \end{aligned}$$

with

$$\begin{aligned} B_1 &= \lambda(t_1)\lambda(t_2)(1-\lambda(t_{(1)}))(1-\lambda(t_{(2)})), \\ B_2 &= \lambda(t_{(1)})\lambda(t_{(2)})(1-\lambda(t_1))(1-\lambda(t_2)). \end{aligned}$$

Here, $\mu = 1$ if $|\hat{\eta}_1| > |\hat{\eta}_2|$ and 0 otherwise, and $\delta = 1$ if the pretest procedure is general-to-specific and 0 if the procedure is specific-to-general.

Because E(UR) depends on 5 parameters, only a 6-dimensional plot would do full justice to its behavior. This task being beyond us, let us first consider the mean squared error $R = \mathrm{MSE}(W\hat{\eta})$ and the expected reported variance $\mathrm{E}(W)$ for the two procedures. Both functions depend on η_1, η_2, and r. The $\mathrm{E}(W)$ is always bounded, as noted in Section 4. The matrix R is also bounded in the general-to-specific procedure, but R can be unbounded in the specific-to-general procedure. More specifically,

$$\max_{\eta_1,\eta_2} R(\eta_1,\eta_2,r) \to \infty \qquad \text{as } r \to 1,$$

when the procedure is specific-to-general. This very different behavior of R in the two procedures is reflected in Figure 7, where we consider

$$\mathrm{E}^{**}(\mathrm{UR}) = \max_{\eta_1,\eta_2} \mathrm{E}^{*}(\mathrm{UR}) = 1 - \min_{\eta_1,\eta_2} \min_{1\le j\le m} \xi_j(R^{-1/2}(\mathrm{E}\,W)R^{-1/2}),$$

as a function of r.

FIGURE 7

For both procedures the function $\mathrm{E}^{**}(\mathrm{UR})$ is symmetric around $r = 0$. For $r = 0$ the two procedures are the same and the function value is almost 0.90. In the specific-to-general procedure, $\mathrm{E}^{**}(\mathrm{UR})$ increases monotonically to 1 as r increases from 0 to 1. The general-to-specific procedure has a uniformly lower $\mathrm{E}^{**}(\mathrm{UR})$, its behavior is non-monotonic, and it converges to 0.87 as $r \to 1$, the same maximum value as in the case $m = 1$ (depicted as a horizontal line in the figure). The difference between the two procedures is especially large when r is close to 1, that is when Mz_1 and Mz_2 are strongly correlated. This can be understood as follows. Let $r = 1$ and let $\eta_1 = -\eta_2 = \bar{\eta}$, say. Then, for large $\bar{\eta}$, the probability of choosing one of the partially restricted models $\mathcal{M}_1$ or $\mathcal{M}_2$ approaches 0. In the specific-to-general case, we will choose the restricted model $\mathcal{M}_0$ with probability approaching 0.95 and model $\mathcal{M}_{12}$ with probability approaching 0.05. Hence, for $r = 1$ and $\bar{\eta} \to \infty$, we find that E(UR) approaches 1 for any q_0^2. (In fact, the MSE of the pretest estimator is unbounded and proportional to $\bar{\eta}^2$ when $\bar{\eta}$ approaches ∞.) But in the general-to-specific case, the MSE is always bounded and hence $\mathrm{E}^{*}(\mathrm{UR}) < 1$, using (8).

Although the functions are continuous, there are various kinks. This is the result of the fact that there exist various local maxima. At a kink we move from one local maximum to another local maximum. Clearly, underreporting can be a very serious problem and, for $m \ge 2$, can be essentially unbounded, depending on the chosen pretest procedure.

For $r = 0$ the worst case gives $\mathrm{E}^{**}(\mathrm{UR}) = 0.87$ for $m = 1$ and 0.90 for $m = 2$. We now ask how underreporting depends on m. There are 2^m models to consider and one may think therefore that 'badness' increases by a factor of 2^m. On the other hand, all t-statistics are functions of only m random variables $\hat{\eta}_1, \ldots, \hat{\eta}_m$, so that 'badness' increases possibly only by a factor of m. We consider the special case where $Z'MZ = I_m$. Then all vectors

Mz_i are orthogonal, and the m-dimensional problem collapses in to m one-dimensional problems (Theorem 3). All pretest procedures are the same in this case, and the maximum E**(UR) is plotted in Figure 8 as a function of m.

FIGURE 8

The figure reveals that E**(UR) increases with m but less than linearly. In fact, we find that the actual pretest mean squared error is about $7.3m^{0.45}$ times the expected reported variance when $1 \leq m \leq 5$ and about $4.5m^{0.76}$ when $m > 6$. Although this result is valid only when $Z'MZ = I_m$, it nevertheless suggests that the increase in 'badness' is not as fast as one might have feared.

In a practical situation, we know q_0^2, q, and r, but not η_1 and η_2. Let us analyze one such situation where $q_0^2 = 2$, $q = (1/3, (2/3)\sqrt{2})'$ (so that $q'q = 1$), and $r = 0.8$.

FIGURES 9 AND 10

Figures 9 and 10 give the E(UR) as a function of η_1 and η_2, first for the general-to-specific procedure, then for the specific-to-general procedure. The E(UR) lies always between 0 and 1, and is symmetric around the point $(\eta_1, \eta_2) = (0, 0)$. The functional dependence on (η_1, η_2) is quite complicated, and also quite different for the two procedures. In the general-to-specific procedure (Figure 9), E(UR) is 0 at $(\eta_1, \eta_2) = (4, -4)$, but can be as large as 0.6551 at $(0.4, 1.6)$. In the specific-to-general procedure (Figure 10), E(UR) varies from around 0 at $(4, 4)$ to 0.8798 around the point $(4, -4)$. In this case (and in general), the specific-to-general is more sensitive to underreporting than the general-to-specific procedure.

The contours in the (η_1, η_2) plane are iso-value curves: the darker (redder) the line, the higher the value.

Now consider a specific point $(\eta_1, \eta_2) = (1, -1)$. In Figure 11, we ask what happens in the 6-dimensional picture if we change the five parameters η_1, η_2, q_0^2, q_1, and r, one at a time.

FIGURE 11

At the chosen point, for both procedures, the E(UR) is an increasing function of q_0^2 (and q_2), but decreasing in η_1, η_2, q_1, and r. Figure 11 confirms that the

E(UR) depends strongly, and not symmetrically, on η_1 and η_2. We already know that E(UR) is an increasing function of q_0^2, but the dependence is much less strong for the general-to-specific procedure than for the specific-to-general procedure. The E(UR) also depends strongly on q (that is q_1). Hence, different linear combinations of the β-parameters are affected differently by the pretest procedure. Sensitivity plots like Figure 11 can thus be used to assess the dependence of the E(UR) on the unknown parameters η_1 and η_2, and also on possible measurement error in the observed quantities q_0^2, q, and r.

8 Extensions and conclusions

In this paper we have analyzed the effect of ignoring the model selection procedure in reporting the bias and variance of the commonly used least-squares estimator. We conclude that underreporting is a very serious problem and that not reporting the correct pretest bias and variance can lead to very misleading results. The pretest bias appears to be less of a problem that the pretest variance.

When we have m auxiliary regressors $z_1, \ldots, z_m$, there are 2^m models to choose between. There are many different possible (viable) procedures to select the model. We find that the choice of model selection procedure (for example, general-to-specific or specific-to-general) matters a lot, and that the general-to-specific procedure seems to have more desirable properties. The influence of the selection procedure is higher when the correlation between the z variables (measured by $Z'MZ$) is high, than when it is low. If we can choose the auxiliary regressors such that they are 'orthogonal' (that is, $Z'MZ = I_m$), then all pretest procedures are the same, and hence the sampling properties of the estimators do not depend on the model selection procedure.

As the number of auxiliary regressors m grows, the dangers of underreporting grow as well, but less than linearly, in the sense that the MSE of the pretest estimator is approximately Am^α times the expected reported variance for some $0 < \alpha < 1$.

The paper shows not only that ignoring model selection can lead to serious underreporting, but also provides explicit formulae to calculate the correct

bias, variance, and mean squared error, which are easy to implement in standard packages.

We now discuss briefly three extensions of the results obtained so far.

Unknown σ^2. Although Theorems 1 and 2 are valid whether or not σ^2 is known, the rest of the paper assumes that σ^2 is known. This is of course unrealistic and we need to address the question how the results are affected when σ^2 is unknown. As an example, let us consider the case of Figure 3 where $m = 1$, $q_0^2 = \infty$ and $c = 1.96$. When σ^2 is known, the E(UR) takes the values 0.82, 0.86, 0.79, and 0.19 for η equal to 0, 1, 2 and 4 respectively. When σ^2 is not known the calculations are more involved and depend on the degrees of freedom $n - k - m$. The results are summarized in Table 1.

	η			
$n-k-m$	0	1	2	4
10	0.76	0.83	0.77	0.26
30	0.80	0.85	0.78	0.22
50	0.81	0.86	0.79	0.21
∞	0.82	0.86	0.79	0.19

Table 1. E(UR) as a function of the d.f. $n - k - m$ (σ^2 unknown).

We see that the effects of estimating σ^2 are relatively small, especially in the region of interest where $|\eta|$ is around 1 or 2. Although this example is typical for the behavior of the E(UR), more work is needed in this direction, especially for $m \geq 2$.

Misspecification. We have also assumed that the unrestricted model is the data-generating process. Again, this may not be realistic, and we shall consider what happens if in fact a larger model generates the data. Thus, we assume that the data-generating process is $y = X\beta + Z_1\gamma_1 + Z_2\gamma_2 + \varepsilon$, but that we have no data on Z_2. The Equivalence Theorem is still applicable in this situation. Since Z_2 is not known, the model selection takes place under the restriction $\gamma_2 = 0$. The bias of the pretest estimator b will be affected by this type of misspecification, but not the variance. Under the simplifying assumption that $Z_1'MZ_2 = 0$, and letting $Q_i = (X'X)^{-1}X'Z_i(Z_i'MZ_i)^{-1/2}$

$(i = 1, 2)$, we have

$$\mathrm{E}(\omega' b) = \omega'\beta - \sigma\left(\omega' Q_1 \,\mathrm{E}(W_1\hat{\eta}_1 - \eta_1) - \omega' Q_2 \eta_2\right),$$

and hence the misspecification has an effect on the bias and the mean squared error of $\omega' b$ only through the scalar $\omega' Q_2 \eta_2$, which of course is unknown. Notice that the bias (in absolute value) and the mean squared error can either decrease or increase because of the misspecification.

Asymptotics. Since all results in the paper are exact finite-sample results, we now ask how the estimators behave for $n \to \infty$. We assume that $n^{-1}(X : Z)'(X : Z)$ approaches a finite positive definite limit. Theorem 2 implies that

$$\mathrm{E}\, b = \beta - \sigma(n^{1/2}Q)\,\mathrm{E}\left(W(n^{-1/2}\hat{\eta}) - n^{-1/2}\eta\right),$$
$$\mathrm{var}(b) = \frac{\sigma^2}{n}\left(\left(\frac{X'X}{n}\right)^{-1} + (n^{1/2}Q)\,\mathrm{var}(W\hat{\eta})(n^{1/2}Q)'\right).$$

Since W is a weighted average of a finite number of idempotent matrices and $\mathrm{var}(\hat{\eta}) = I_m$, it follows that $\mathrm{var}(W\hat{\eta})$ remains bounded as $n \to \infty$. Hence b is consistent when $n^{-1/2}\left(\mathrm{E}\left(W\hat{\eta} - \eta\right)\right) \to 0$ or, equivalently, when $\mathrm{E}\left(W\hat{\gamma} - \gamma\right) \to 0$.

Following Pötscher (1991, p.164) we shall say that a pretest *procedure* is *(strongly) consistent* if asymptotically the correct 'minimal' model is selected. In general, pretest procedures are not strongly consistent. We shall say that a pretest procedure is *weakly consistent* if, for any $\gamma_i \neq 0$ $(i = 1, \ldots, m)$, the probability that the procedure selects a model without γ_i approaches 0 as $n \to \infty$. Thus, a weakly consistent procedure does not necessarily exclude γ_i from the model when $\gamma_i = 0$, and we may therefore end up with a model that is too large, but not with one that is too small. All the usual common-sense pretest procedures are weakly consistent.[5] It is easy to see that a weakly consistent procedure leads to a consistent estimator (Pötscher 1991, Lemma 2). For simplicity, let $m = 1$. If $\gamma = 0$, then both b_r and b_u are unbiased and

[5]The non-viable procedure defined in footnote 4 is not weakly consistent, because for $\eta \neq 0$, $\lambda \xrightarrow{p} 0$ as $n \to \infty$, implying that $\operatorname{plim} b = \operatorname{plim} b_r \neq \beta$. A viable procedure may or may not be weakly consistent, and a weakly consistent procedure may or may not be viable.

consistent, and hence any weighted average $b = \lambda b_u + (1-\lambda)b_r$ is consistent as well. If $\gamma \neq 0$, then b_u is consistent but b_r is not. However, since the procedure is weakly consistent, $\lambda \xrightarrow{p} 1$, and hence b is consistent.

Given consistency, the variance of the asymptotic distribution of $n^{1/2}(b-\beta)$ follows from the limit of $\text{var}(W\hat{\eta})$ as $n \to \infty$. Again, let $m = 1$. If $\gamma \neq 0$, then $\lambda \xrightarrow{p} 1$ and $\text{var}(\lambda\hat{\eta}) \to \text{var}(\hat{\eta}) = 1$. In this case $\lambda\hat{\eta} - \eta$ converges to a normal distribution with mean zero and variance one. In contrast, if $\gamma = 0$, then $\hat{\eta} \sim \text{N}(0,1)$ and hence, for $c = 1.96$, $\lambda = 1$ with probability 0.05, $\lambda = 0$ with probability 0.95, and $\text{var}(\lambda\hat{\eta}) = 0.28$ (see Figure 1). In this case $\lambda\hat{\eta} - \eta$ converges to a distribution with mean zero and variance 0.28, but this distribution is not normal. See also Leeb and Pötscher (2000), especially equations (28) and (29).

As a result, underreporting may occur even asymptotically. For $m = 1$, underreporting vanishes asymptotically for $\gamma \neq 0$, but not for $\gamma = 0$ when the E(UR) may be as large as $(R(0) - 0.05)/R(0) = 0.82$.

Future work will have to clarify various other issues not covered in this paper. For example, there are selection procedures other than general-to-specific and specific-to-general. How does the E(UR) depend on these? All these pretest procedures are discontinuous; they choose one of 2^m models based on the values of t- and F-statistics. A continuous procedure can also be defined. Theorem 2 allows for this situation, and it will probably lead to better sampling properties; see Magnus(2000) for an analysis of the case $m = 1$.

Also, the E(UR) depends on η which is unknown. What is the best way to estimate the expected underreporting ratio? Simplest is to replace η by $\hat{\eta}$. The properties of this estimator will have to be analyzed. Another possibility is to report the worst possible case, given the observed y, X and Z. That is, to calculate $\max_\eta \text{E(UR)}$. In Figures 9 and 10, for example, the maxima are 0.6551 (general-to-specific) and 0.8798 (specific-to-general), and they give an indication of how bad underreporting can be in a specific situation.

References

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15, 190–204.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society*, Series A, 158, 419–466.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253–265.

Feldstein, M.S. (1973). Multicollinearity and the mean square error of alternative estimators. *Econometrica* 41, 337–346.

Giles, J.A. and D.E.A. Giles (1993). Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys* 7, 145–197.

Giles, J.A. and V. Srivastava (1993). The exact distribution of a least-squares regression coefficient after a preliminary t-test. *Statistics and Probability Letters* 16, 59–64.

Hendry, D.F. (2001). Achievements and challenges in econometric methodology. *Journal of Econometrics* 100, 7–10.

Hoover, K.D. and S.J. Perez (1999). Data mining reconsidered; encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 1–25.

Huntsberger, D.V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics* 26, 734–743.

Judge, G.G. and M.E. Bock (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Judge, G.G. and M.E. Bock (1983). Biased estimation. In Z. Griliches, & M.D. Intriligator (eds.), *Handbook of Econometrics*, Vol. I, Chapter 10, Amsterdam: North-Holland.

Leeb, H. and B.M. Pötscher (2000). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. Discussion paper, Department of Economics, Universität Wien.

Lo, A.W. and A.C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3, 431–467.

Lovell, M.C. (1983). Data mining. *The Review of Economics and Statistics* 65, 1–12.

Magnus, J.R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications* 44, 293–308.

Magnus, J.R. (2000). Estimation of the mean of a univariate normal distribution with known variance. CentER Discussion paper, submitted for publication.

Magnus, J.R. and J. Durbin (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67, 639–643.

Miller, A.J. (1990). *Subset Selection in Regression.* London: Chapman and Hall.

Mittelhammer, R.C. (1984). Restricted least squares, pre-test, OLS and Stein rule estimators: Risk comparisons under model misspecification. *Journal of Econometrics* 25, 151–164.

Mosteller, F. (1948). On pooling data. *Journal of the American Statistical Association* 43, 231–242.

Pötscher, B.M. (1991). Effects of model selection on inference. *Econometric Theory* 7, 163–185.

Pötscher, B.M. and A.J. Novak (1998). The distribution of estimators after model selection: large and small sample results. *Journal of Statistical Computation and Simulation* 60, 19–56.

Roehrig, C.S. (1984). Optimal critical regions for pre-test estimators using a Bayes risk criterion. *Journal of Econometrics* 25, 3–14.

Sclove, S.L., C. Morris, and R. Radhakrishnan (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics* 43, 1481–1490.

Sen, P.K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* 7, 1019–1033.

Thomson, M. and P. Schmidt (1982). A note on the comparison of the mean square error of inequality constrained least squares and other related estimators. *The Review of Economics and Statistics* 64, 174–176.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica* 64, 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrics* 68, 1097–1126.

Zaman, A. (1984). Avoiding model selection by the use of shrinkage techniques. *Journal of Econometrics* 25, 73–85.

Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association* 87, 732–737.

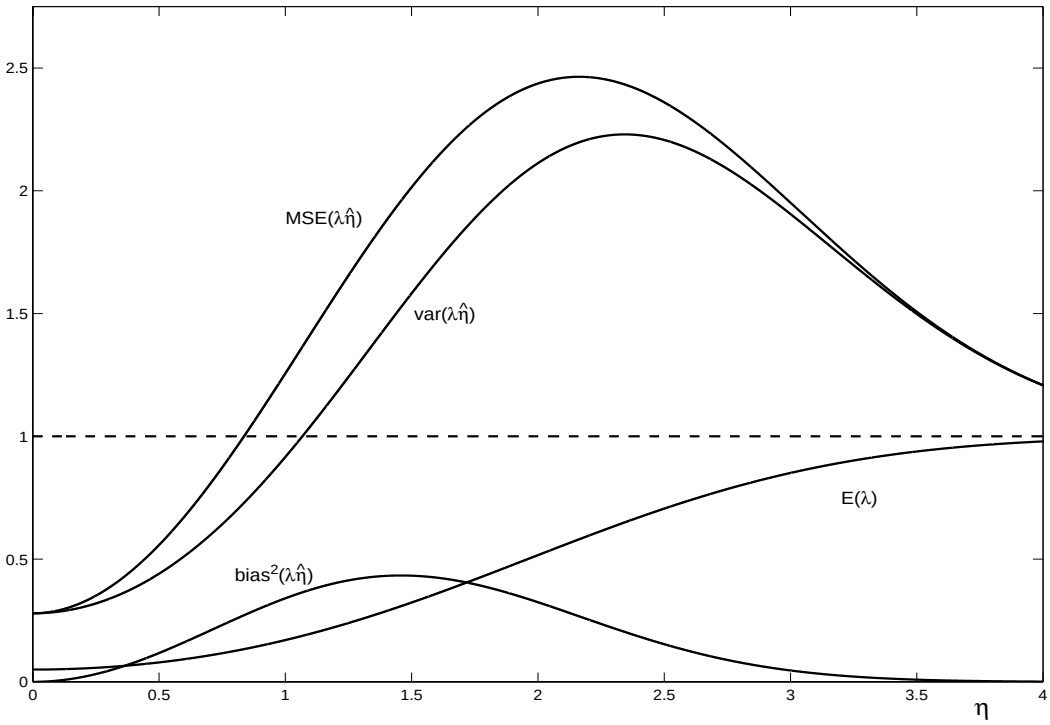

Figure 1. Moments of $\lambda\hat{\eta}$ and λ compared ($m = 1$, $c = 1.96$).

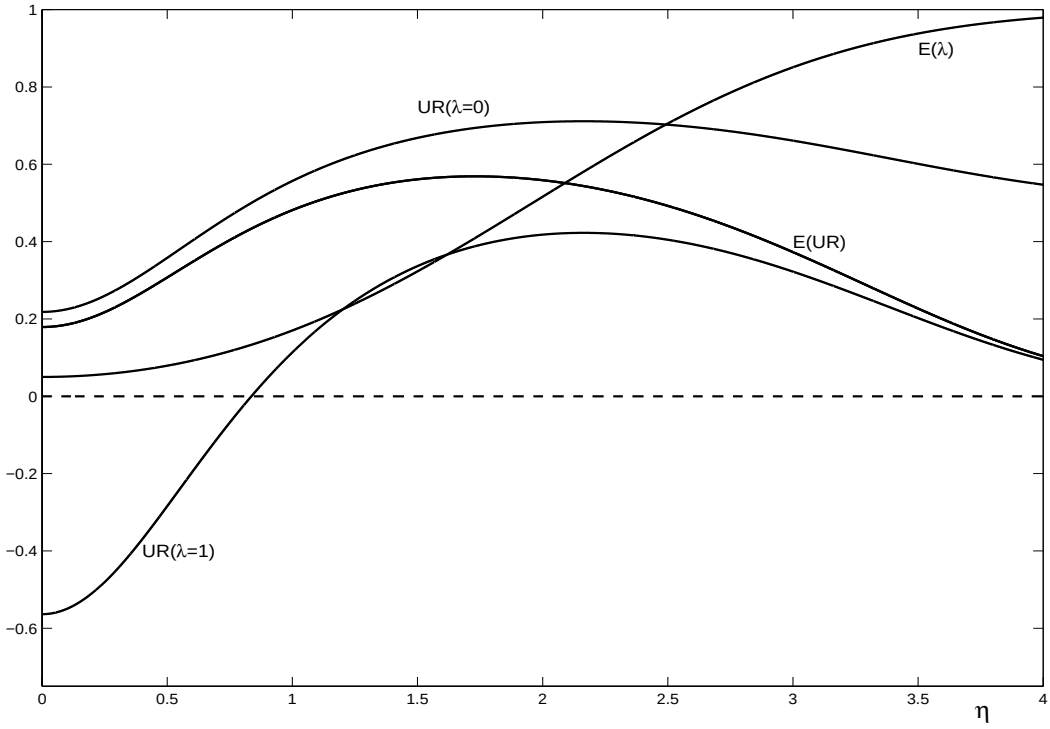

Figure 2. UR (for $\lambda = 0, 1$), E(λ), and E(UR) ($m = 1$, $q_0^2 = 1$, $c = 1.96$).

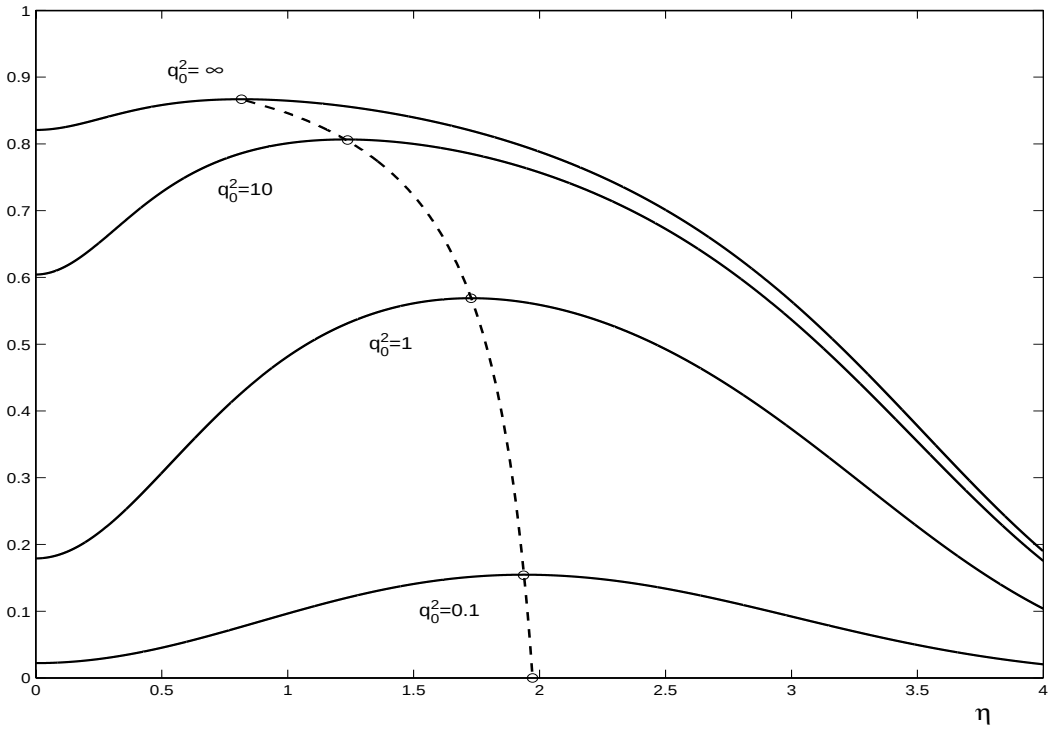

Figure 3. E(UR) and locus of max(E(UR)) ($m = 1$, $c = 1.96$).

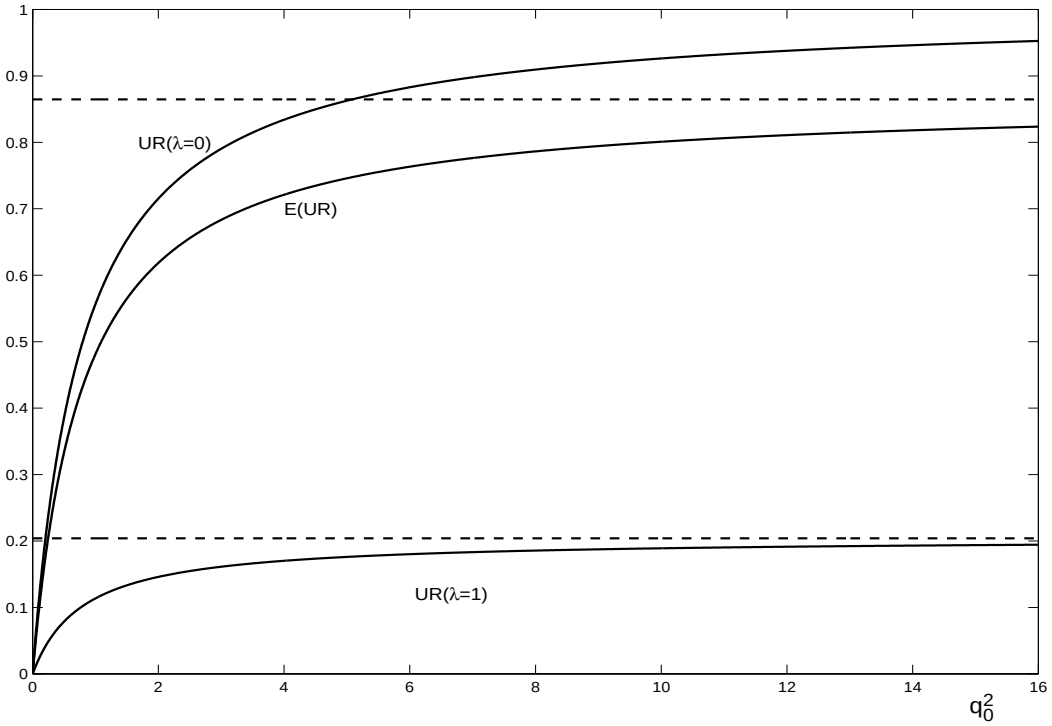

Figure 4. UR and E(UR) as a function of q_0^2 ($m = 1$, $c = 1.96$, $\eta = 1$).

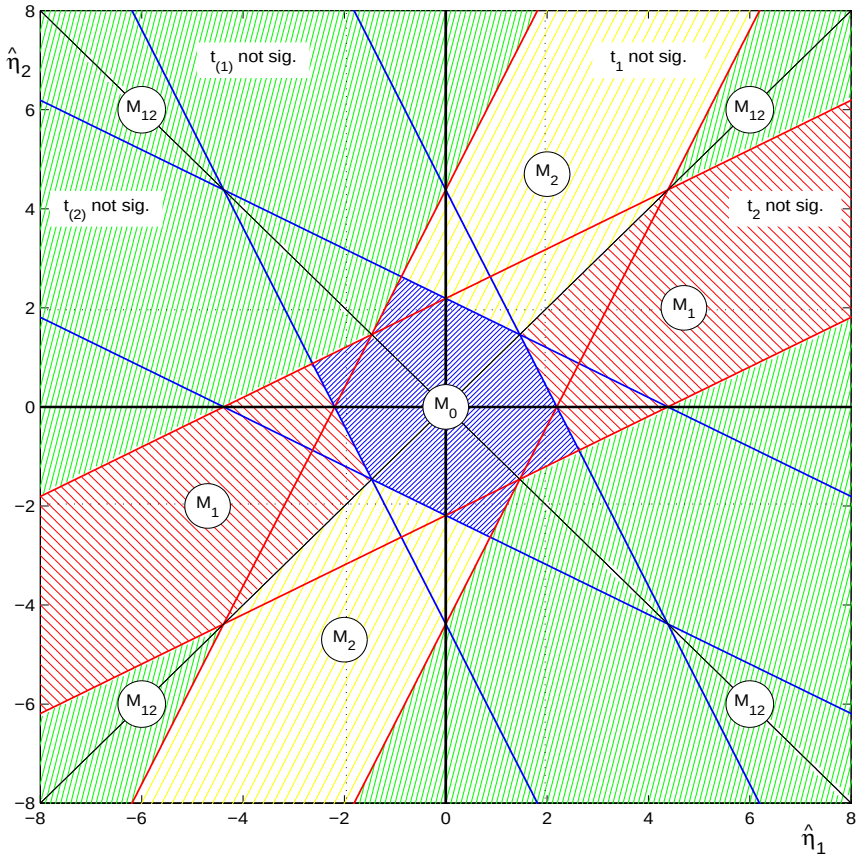

Figure 5. Model selection regions: general-to-specific.

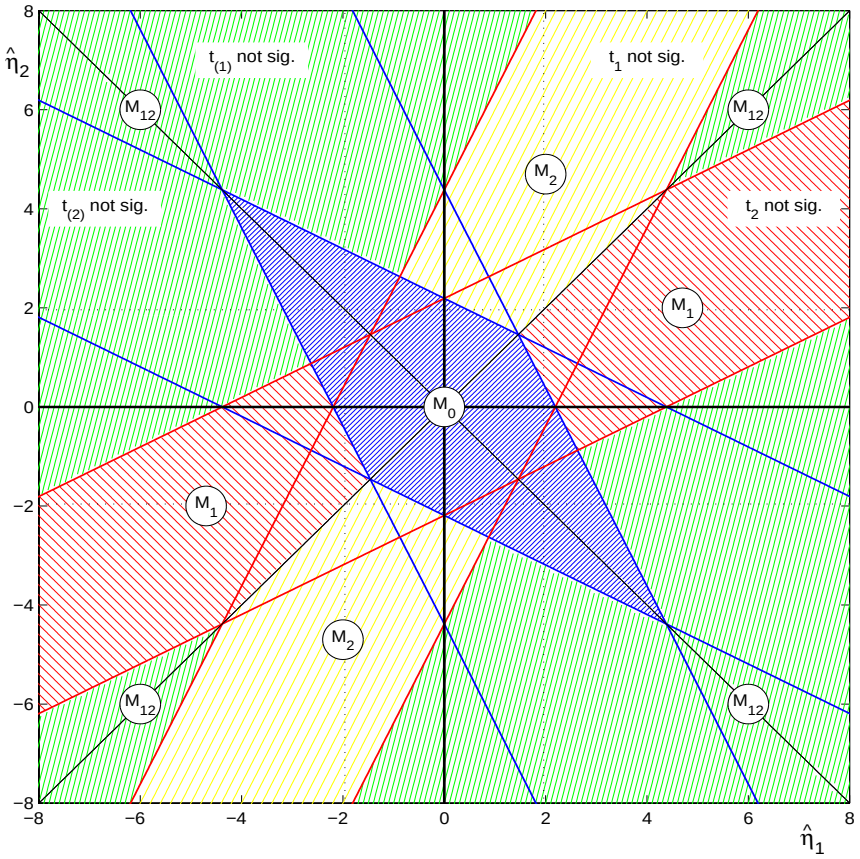

Figure 6. Model selection regions: specific-to-general.

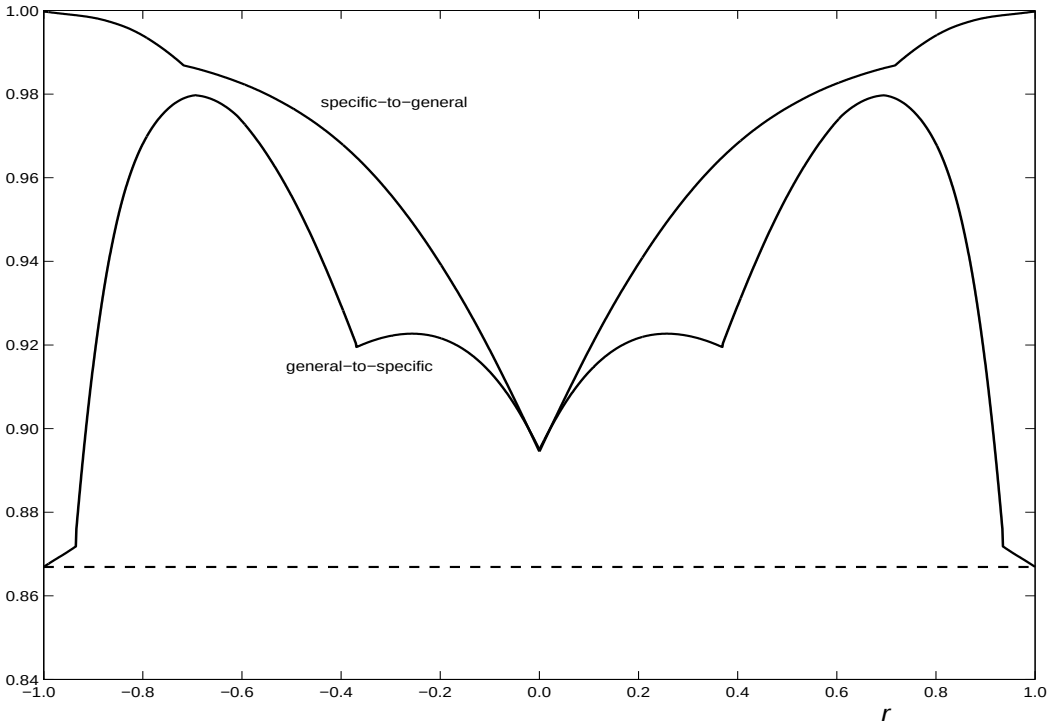

Figure 7. max(E(UR)) as a function of r ($m = 2$).

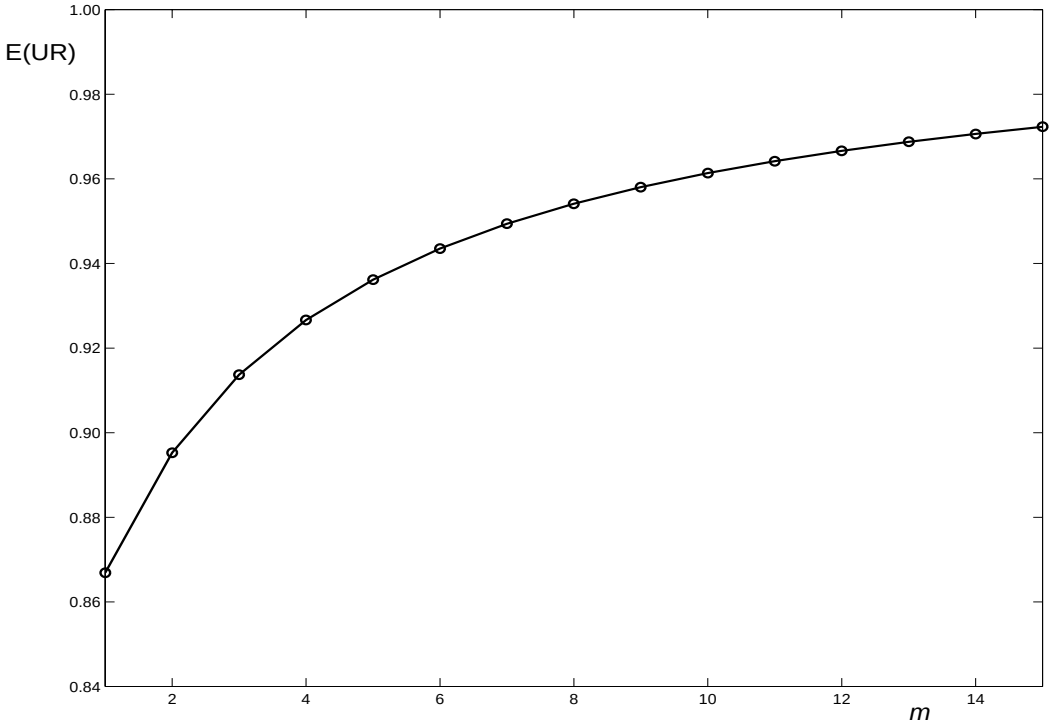

Figure 8. max(E(UR)) as a function of m ($Z'MZ = I_m$).

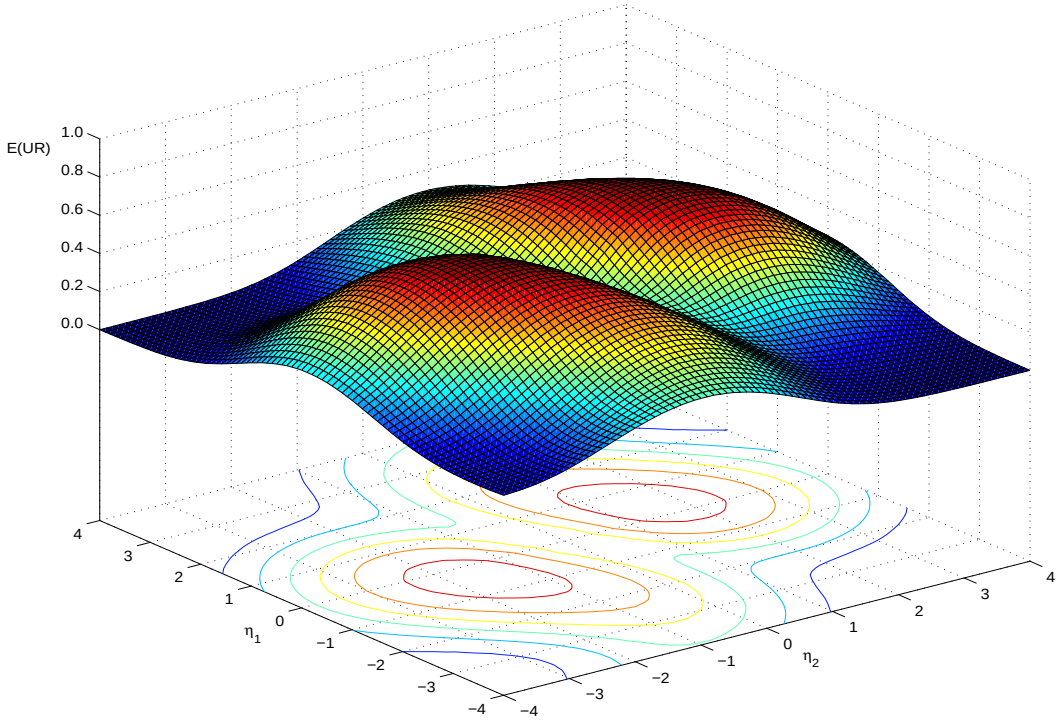

Figure 9. E(UR) as a function of η_1 and η_2: general-to-specific.

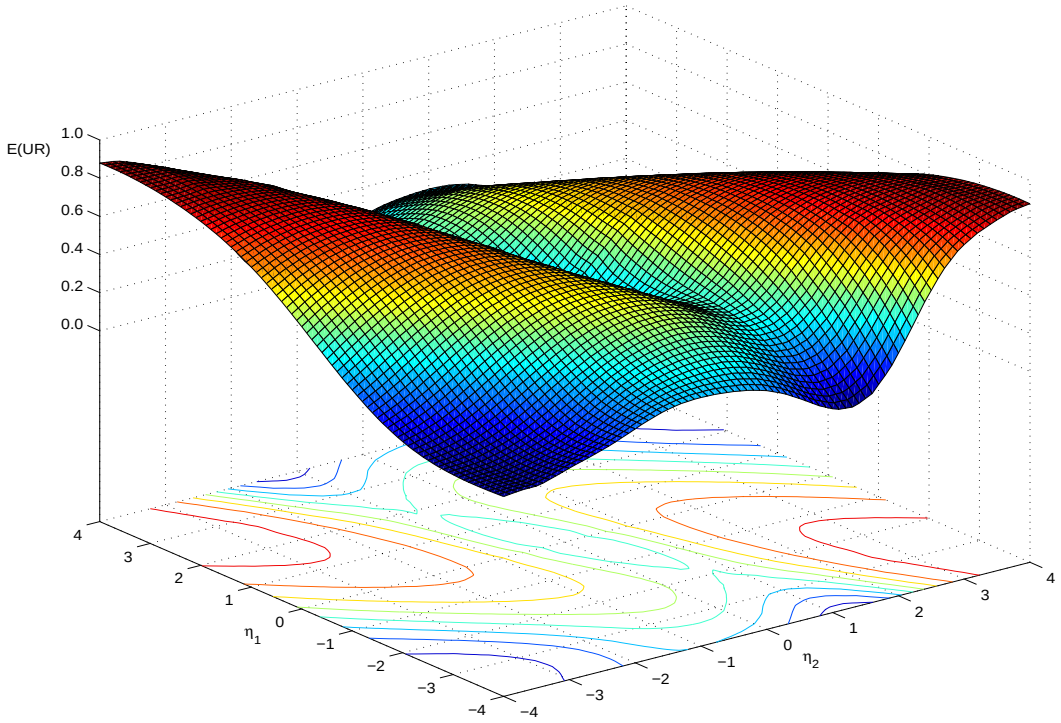

Figure 10. E(UR) as a function of η_1 and η_2: specific-to-general.

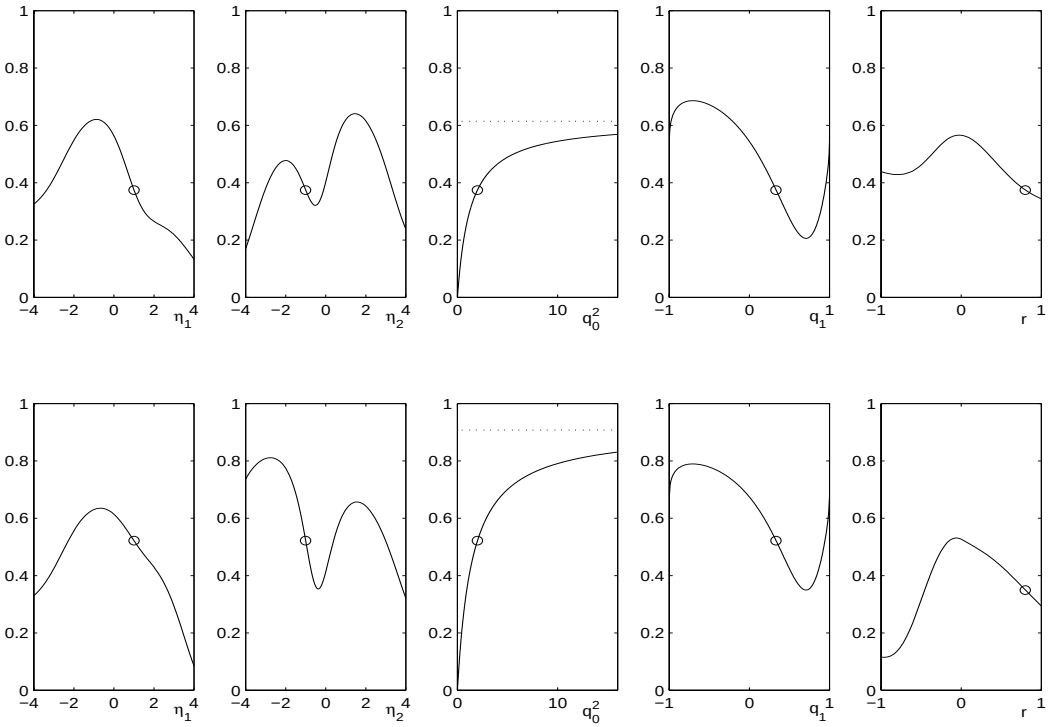

Figure 11. Sensitivity analysis for E(UR):
general-to-specific (top) and specific-to-general (bottom).